CHANGING WHAT'S POSSIBLE

ENLITENED Annual Program Review

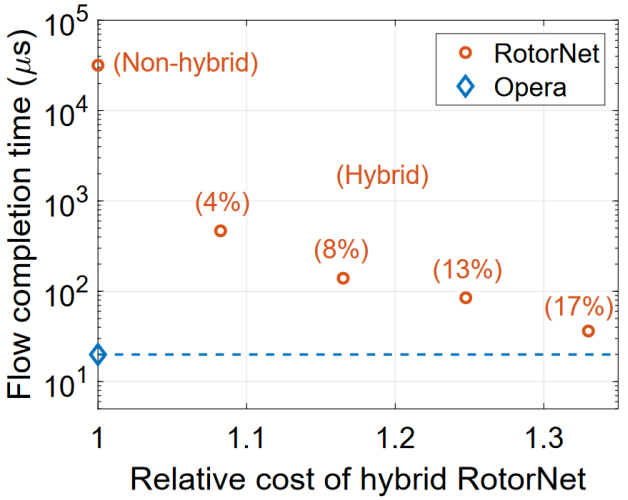
LEED – A Lightwave Energy Efficient Data Center

October 31, 2019

LEED objectives

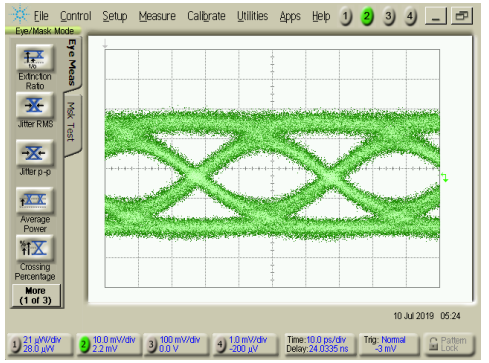
- A robust, scalable, energy-efficient network (ENLITENED Metric 1.1)
- Co-optimized across:
  - Network Architecture
    - Efficient all-optical networks
    - Cost effective and fault tolerant
  - Optical Switch
    - Decouples switching from routing
    - Based on laser-written “pinwheel”
  - Commercially-viable enhanced link-margin interconnects
    - Burst-mode APD receiver
    - WDM modulator array
    - Broadband mux/demux
    - Integrated Tx and Rx (Phase 2)

Efficient all-optical network (Opera)



Racked “Pinwheel” Rotor Switch

25 Gb/s APD for burst-mode Rx

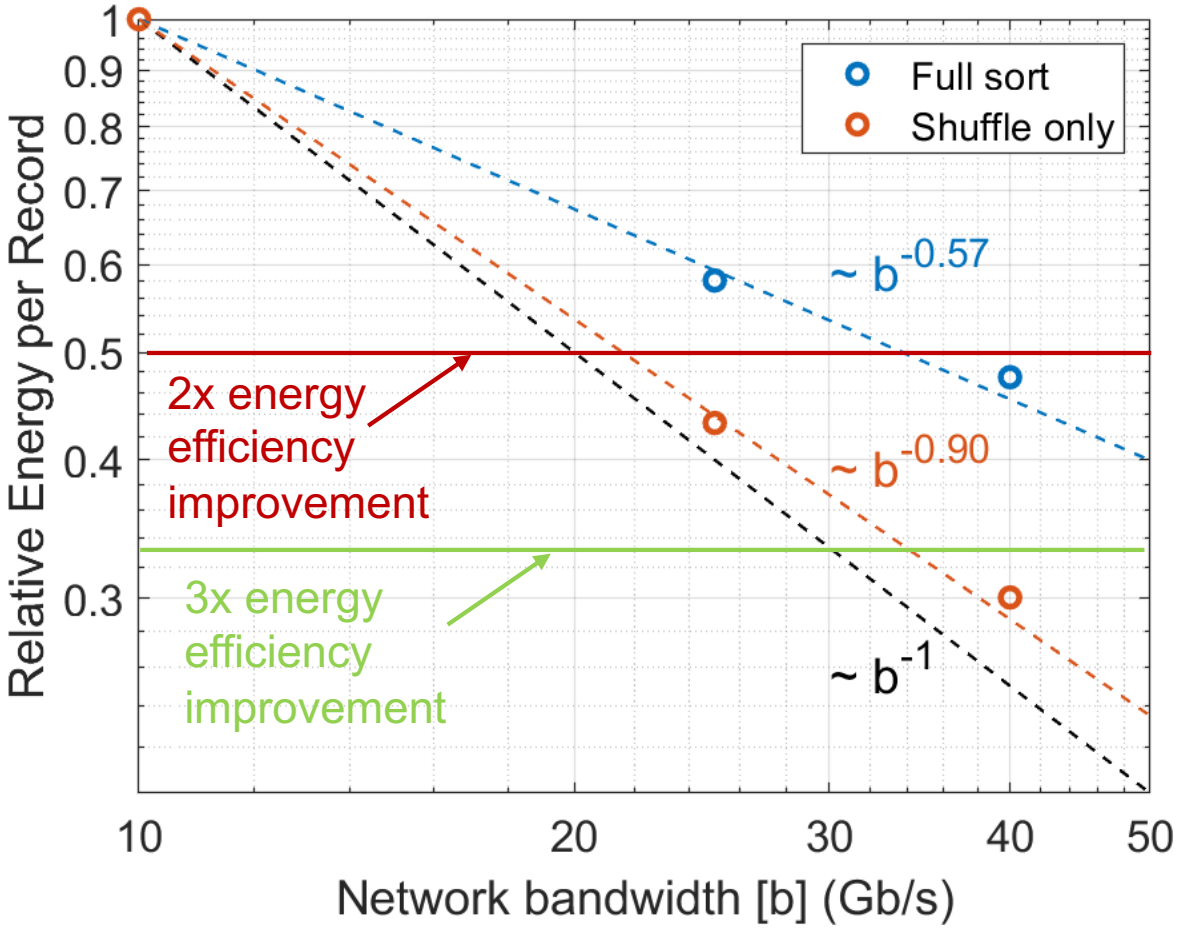


Phase 2 LEED team

- **Networking**
  - George Papen, George Porter, Alex Snoeren (UCSD)
  - Max Mellette (inFocus Networks)
  - Simon Hammond (Sandia National Labs)
- ***Optical Switch***
  - Ilya Agurok, George Papen (UCSD)
  - Joseph Ford, Max Mellette (inFocus Networks)
- ***Interconnects***
  - Shaya Fainman, Shayan Moookherjea (UCSD)
  - Y. Ehrlichman, I. G. Yayla, J. Simons, J. E. Cunningham, A. V. Krishnamoorthy(Axalume)
  - Michael Gehl, Christopher T. DeRose, Paul S. Davids, Douglas C. Trotter, Andrew L. Starbuck Christina M. Dallo, Dana Hood, Andrew Pomerene and Tony Lentine(Sandia National Labs)

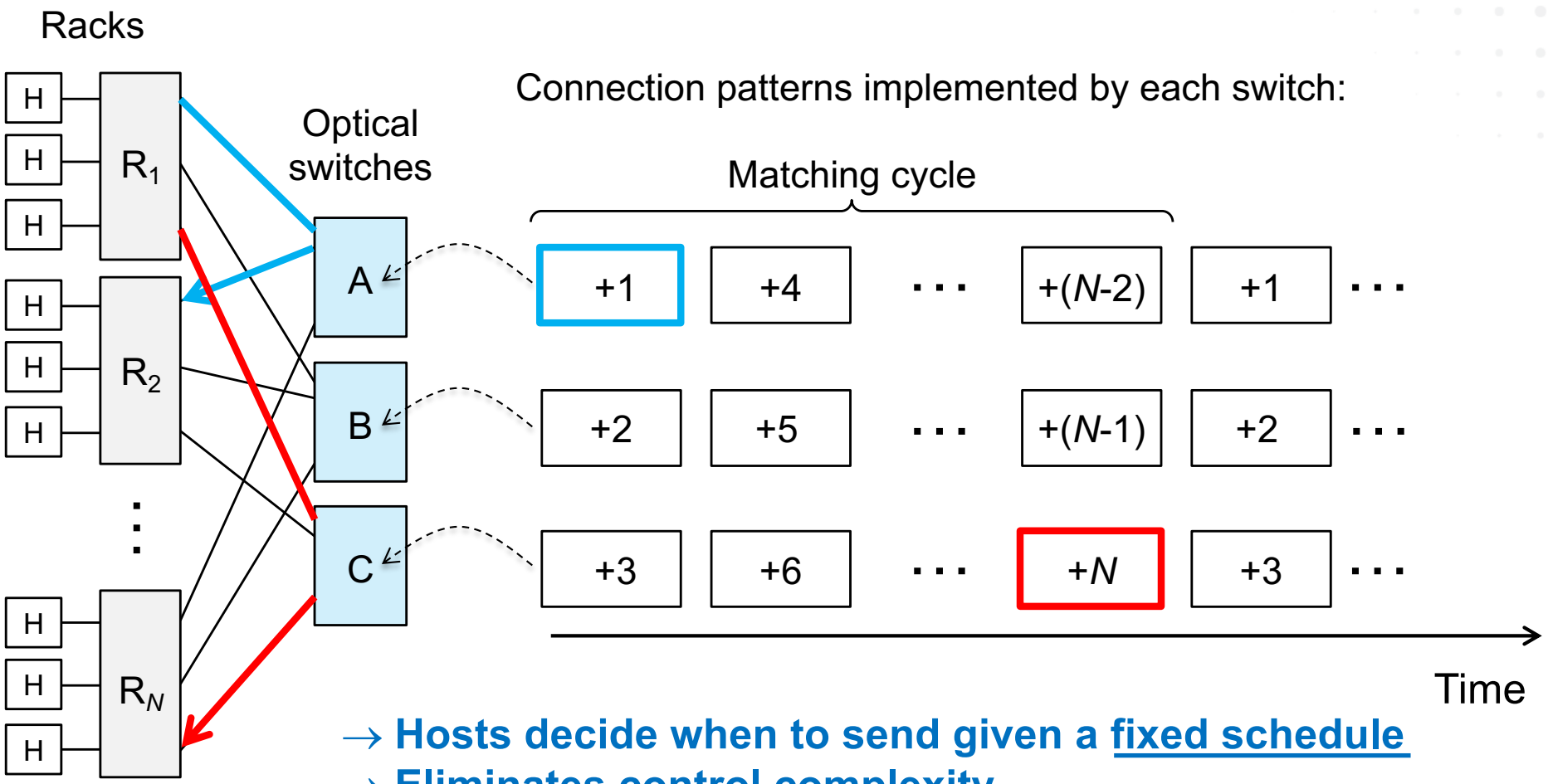
Solving the energy-efficiency challenge: Reduced bandwidth per buck via optical switching

- For shuffle-bound applications:
  ~2x bandwidth → 2x higher energy efficiency
- For sort application:
  ~3x bandwidth → 2x higher energy efficiency
- "Bandwidth per buck" determines actual operating point



Physically measured confirmation of fundamental LEED tenet:
Different applications have different energy-efficiency slope vs bandwidth, but greater network bandwidth leads directly to higher energy efficiency!

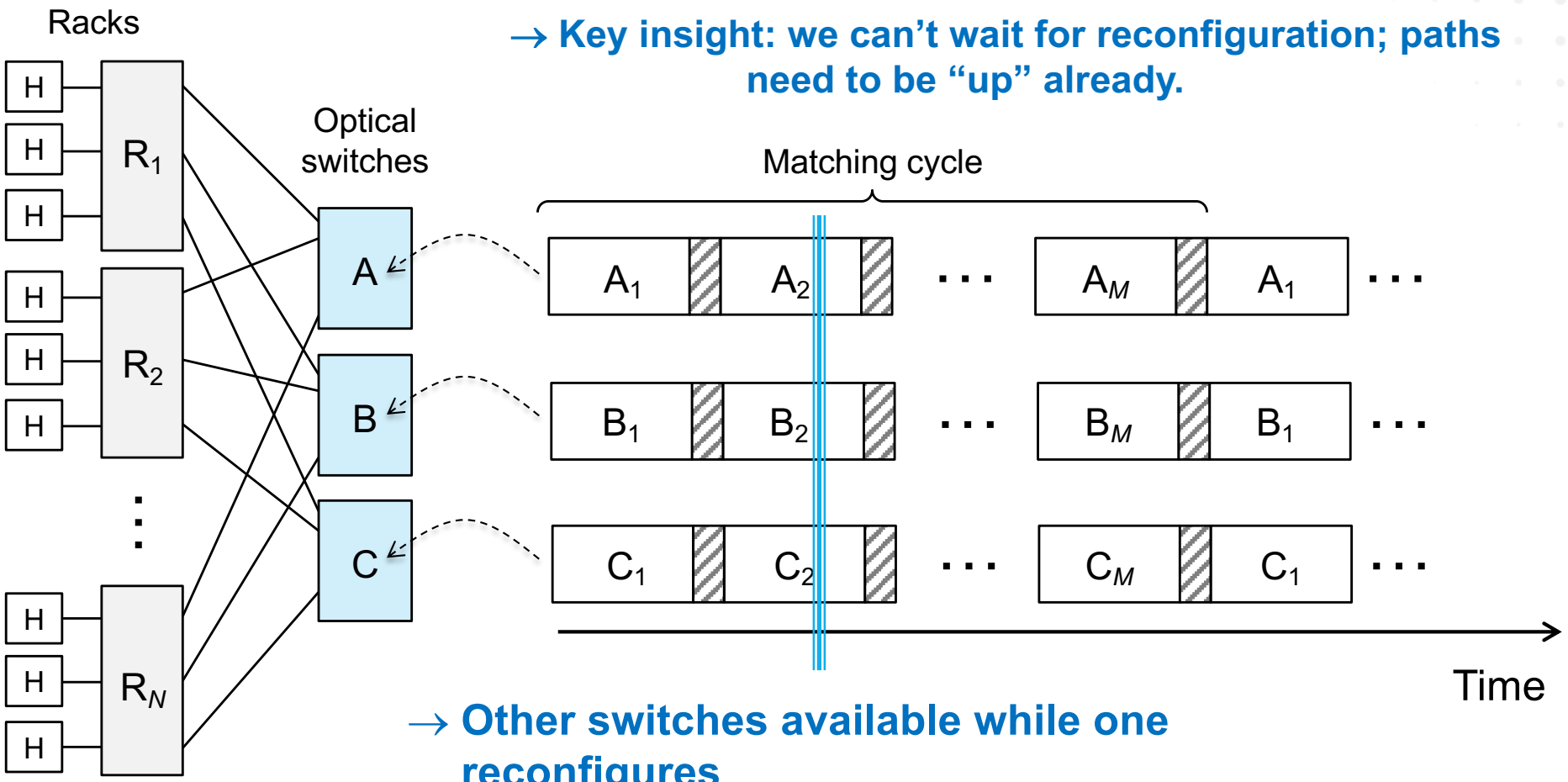
Solving the control plane challenge: RotorNet (Sigcomm '17):



→ **Hosts decide when to send given a fixed schedule**

→ **Eliminates control complexity**

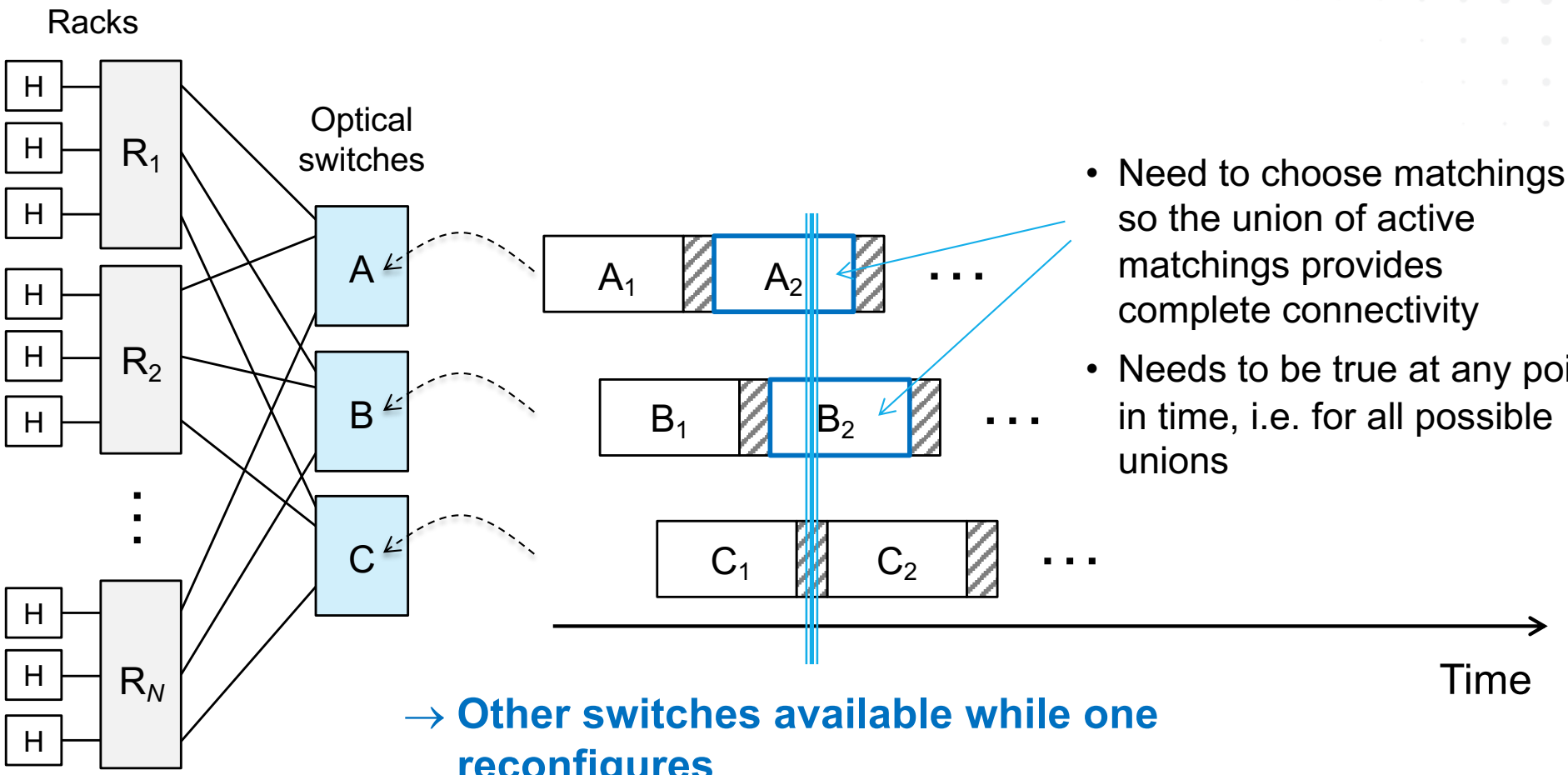
Solving the latency challenge: Opera protocol (NSDI 2020)



→ **Key insight: we can't wait for reconfiguration; paths need to be "up" already.**

→ **Other switches available while one reconfigures**

Solving the latency challenge: Opera protocol (NSDI 2020)

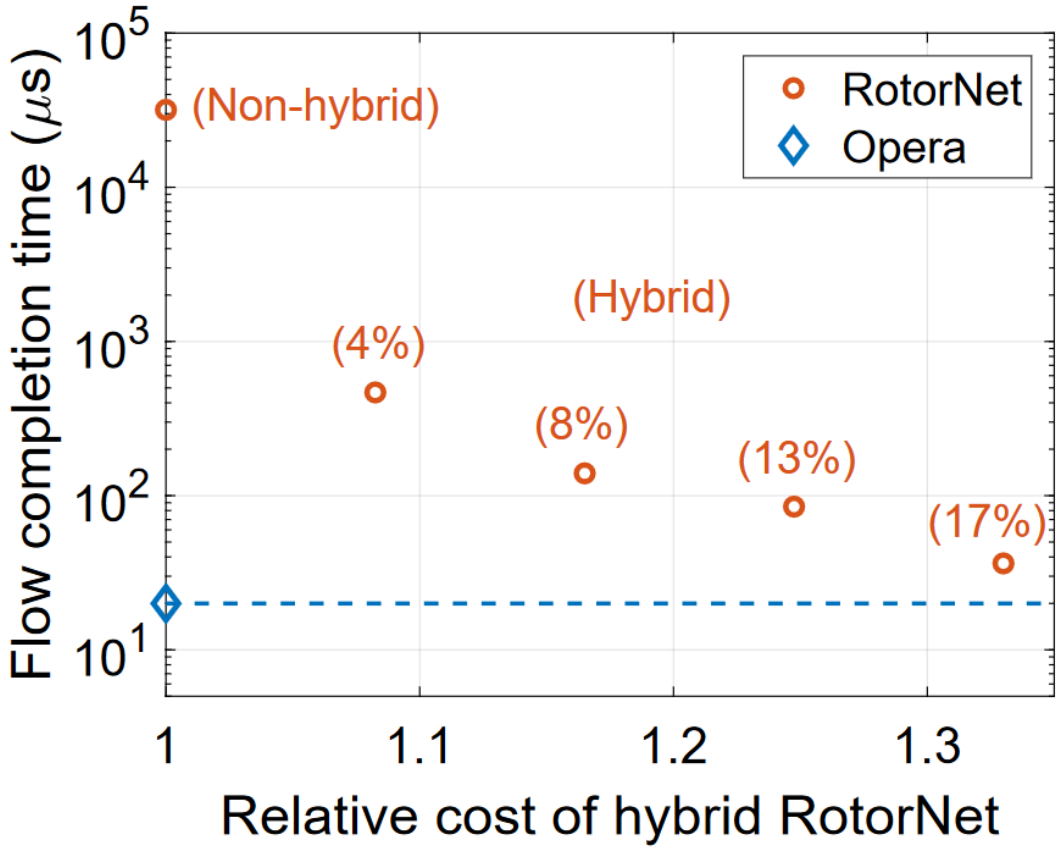


- Need to choose matchings so the union of active matchings provides complete connectivity
- Needs to be true at any point in time, i.e. for all possible unions

→ **Other switches available while one reconfigures**

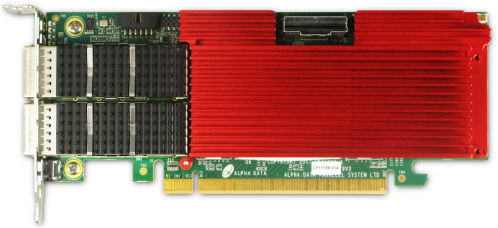
Opera Performance: Out-performs hybrid RotorNet network

- An all-optical transport is *more* efficient than a hybrid packet/circuit network based on RotorNet
- Much more efficient than standard packet-switched network

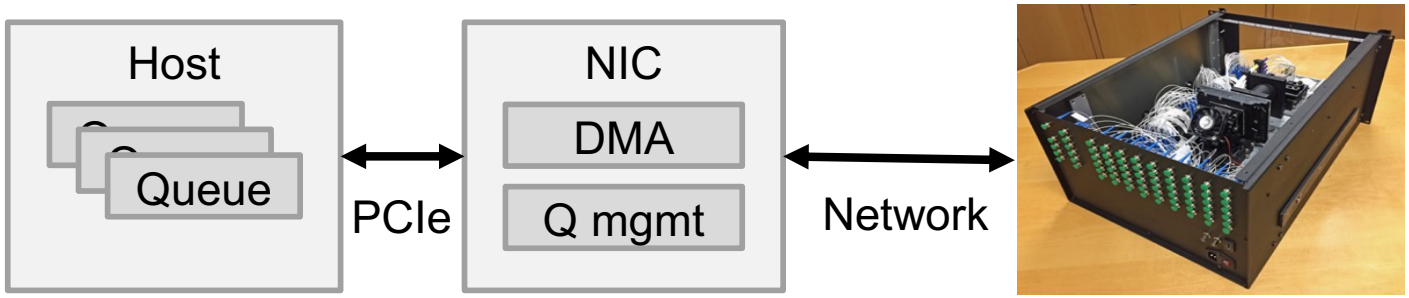


Solving the synchronization challenge: Open-source Corundum FGPA-based NIC

- PCIe interface
  - Provides high-performance (25 Gb/s) Direct Memory Access (DMA) engine



- Software driver
  - Connects software networking stack to FPGA- based NIC & circuits
- Scalable queue management/synchronization
  - 100 ns precision using PTP for TDMA (RotorNet/Opera)
  - 1000+ independent, hardware-managed queues!



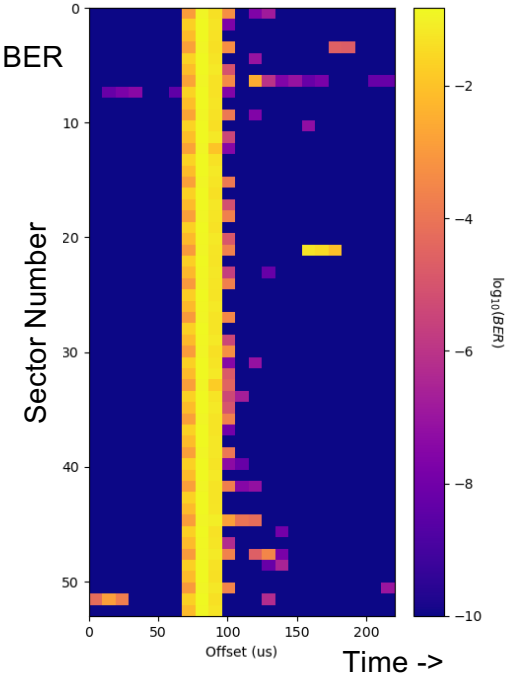
Provides the "glue" between packet and circuit worlds

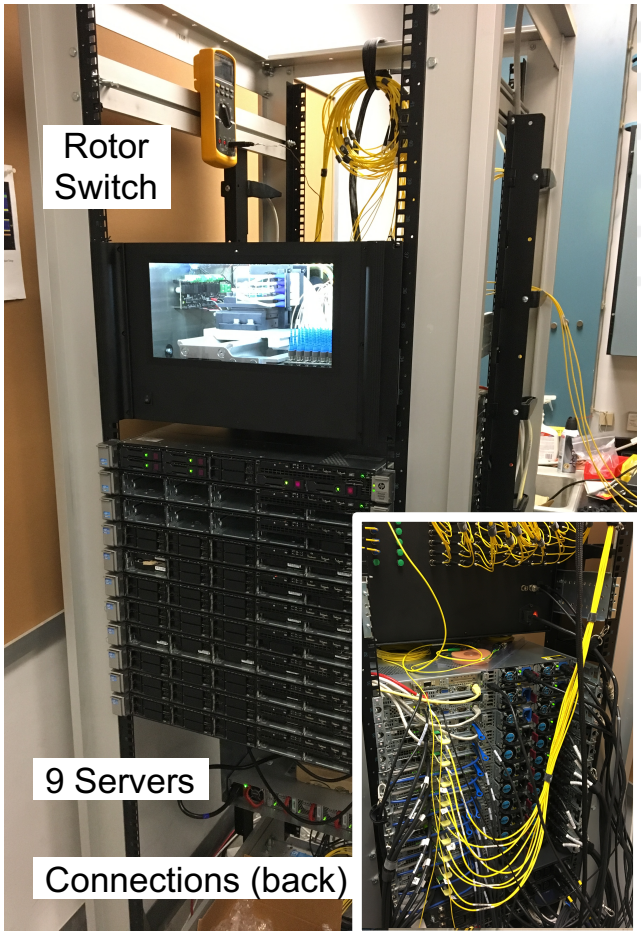
Source code:https://github.com/ucsdsysnet/corundum

Architectures: Phase 1 accomplishments/ Phase 2 Objectives

- Phase 1 accomplishments
  - Control plane: Full stack demonstration of RotorNet
    - Can run unmodified Linux apps on optical network
  - Latency: Developed and simulated Opera
    - Out-performs hybrid circuit/packet networks
  - Synchronization: Corundum FGPA-based NIC
    - Key interface between circuit and packet worlds
- Phase 2 Objectives
  - Full-stack implementation of Opera
  - Realistic assessment of optical networks using community-established benchmarking
  - Research into viable workloads for HPC/ datacenters using benchmarks

Switching: Rotor switch status

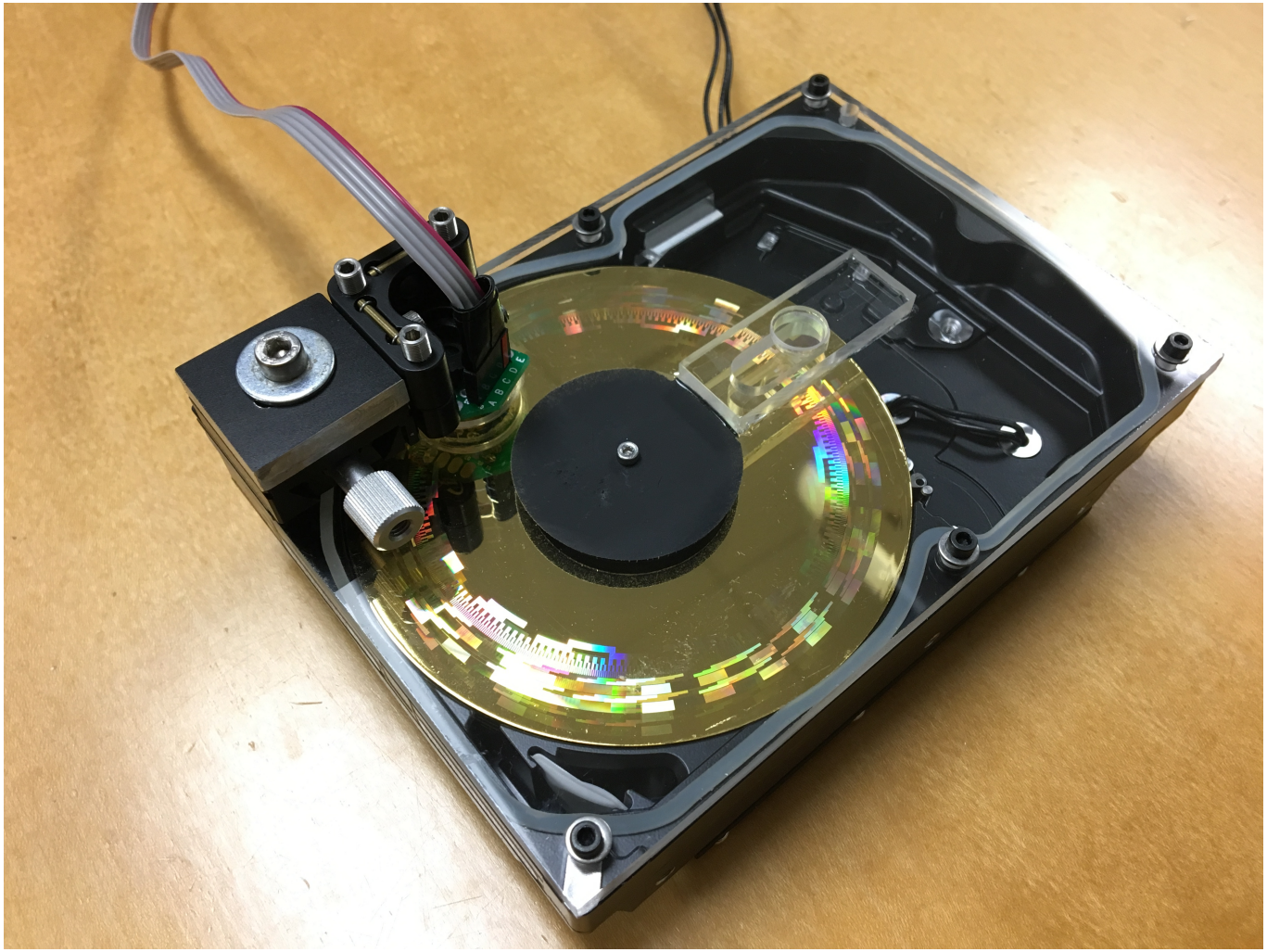




Selector switch integration

Physical format	Rack-mounted	vs “breadboard” spec
Crosstalk	< 30 dB	< 20db requirement)
Operating spectrum	> 120 nm	> 35nm spec
Optical switching time	15 µs	< 75 µs spec
System switching time	40 µs	< 100 µs spec
2-pass insertion loss	5 – 8 dB	< 7 dB spec on average

Prototype pinwheel in 3.5” HGST Deskstar NAS

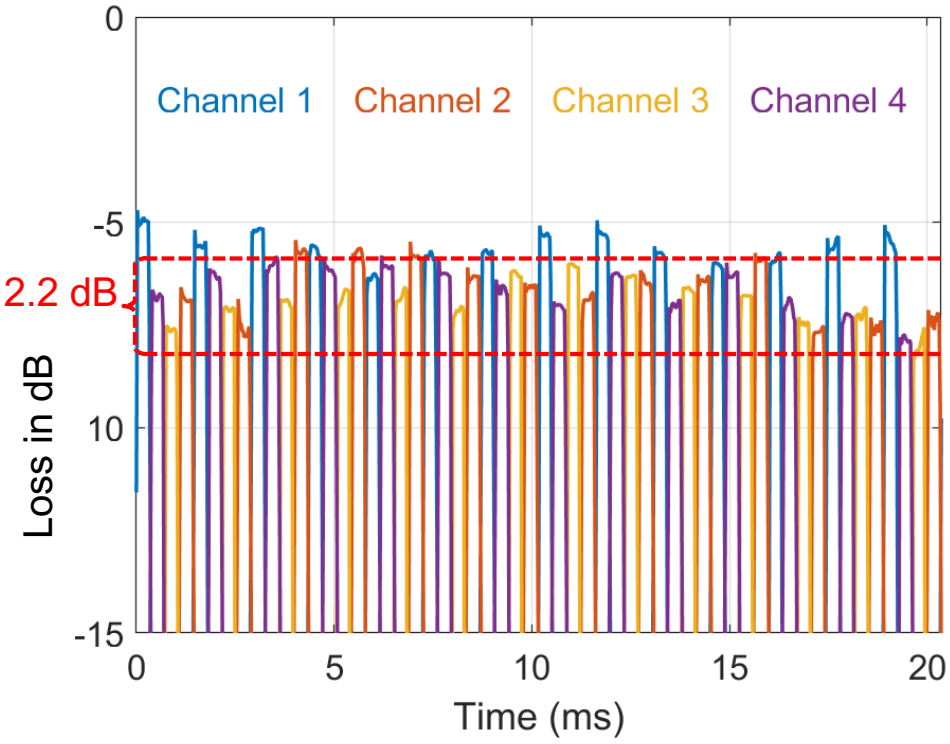


Pinwheel with encoder, encoder tracks, and clear cover

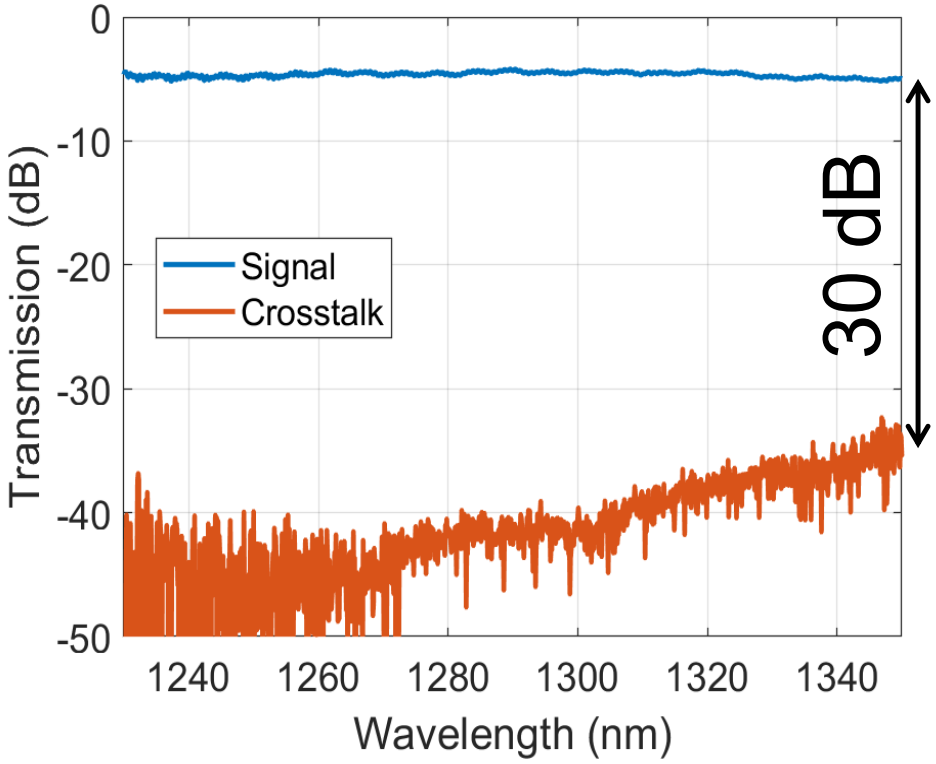
Rotor switch transmission

Doublepass rotor switch insertion loss

5-8 dB overall, up to 2.2 dB on single output

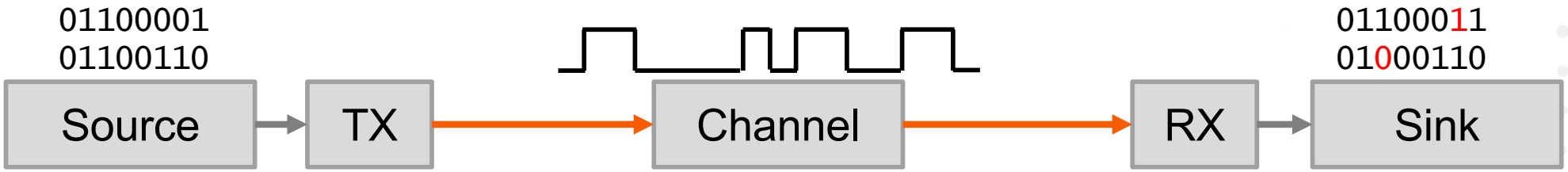


Doublepass crosstalk

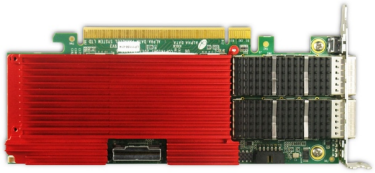
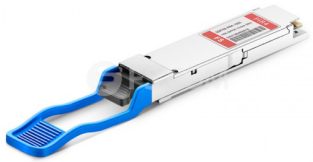
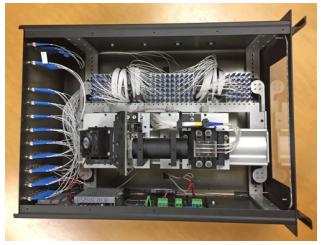


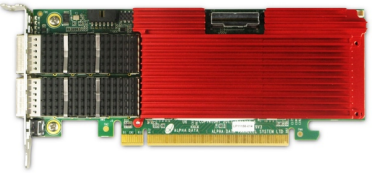
Power variations caused by combination of point defects and stitching errors during fabrication

Switched-network BER measurements



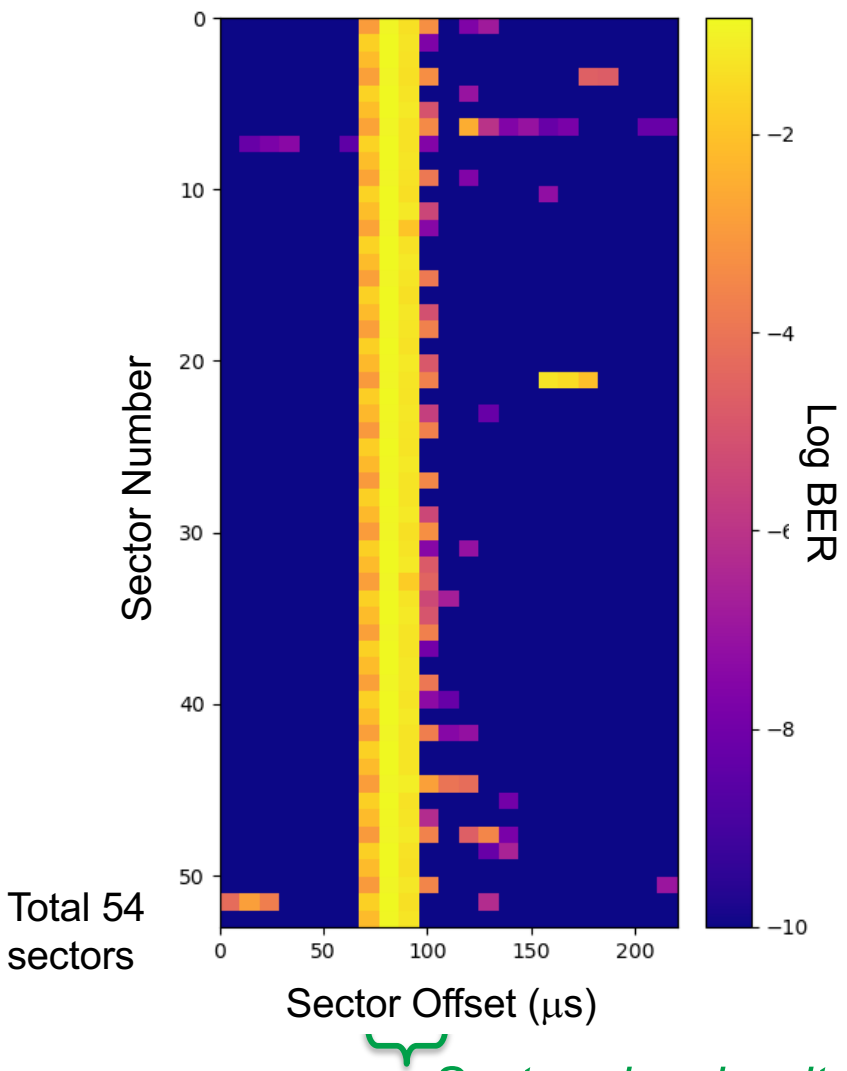
Corundum NIC	COTS TX	RotorNet connection:	COTS RX	Corundum NIC
PRBS31 from a linear feedback chip register	1310nm SMF, TXR, PSM-4	switch / match / switch (unamplified)	1310nm SMF TXR, PSM-4	Reads signal and calculate errors



Automated hardware platform measurement: PRBS generated in Corundum NIC (or FGPA board). Computer used only to control process and retrieve collected heat-map data.

Measured BER "heat map"



Each heat map shows BER vs time for one input connection, through 54 sectors (3 configurations repeated 18 times) of one full disk rotation.

System-level switching time includes:

- Physical switching time (~22 us)
- AGC and CDR lock time (~10 us)
- Disk synchronization (~10 us)

Does not include:

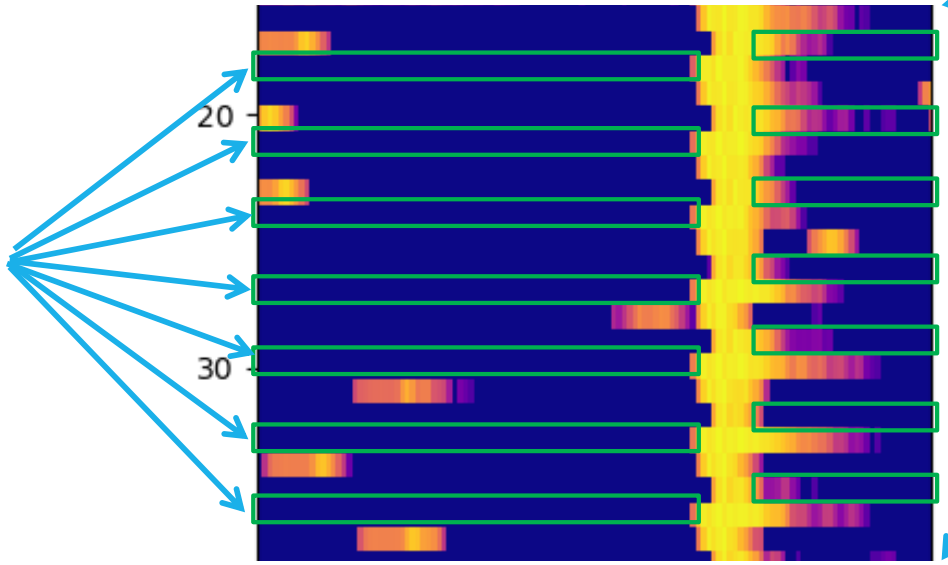
- Ethernet 64b/66b frame sync
- NIC transmit timing accuracy

Demo of full-stack optically-switched network running unmodified Linux app (iperf)

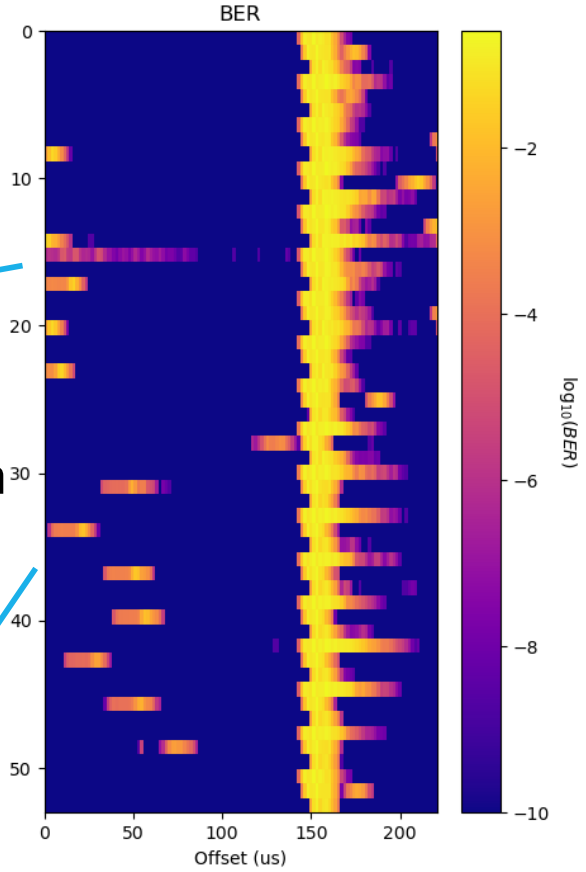
- Structure in BER showed some paths through the rotor switch are usable (with few errors from pinwheel fab errors /power offset)

- Ran app “iperf” on a single path
  - Network measurement application
  - **Unmodified** TCP for network stack

Every third sector has low errors -sufficient to run app.



Zoom in



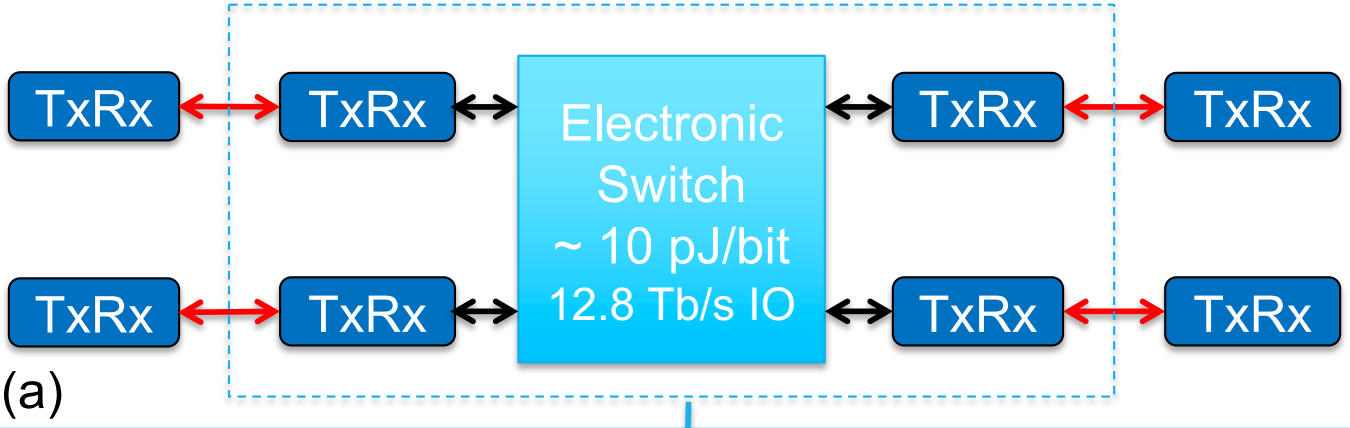
Switches: Phase 1 accomplishments/ Phase 2 Objectives

- Phase 1 accomplishments
  - Design, built, and tested first Rotor switch
    - 2-pass insertion loss: 5 – 8 dB
    - Optical switching time 15 μs; System switching time 40 μs
    - Bandwidth > 120 nm; crosstalk < -30 dB
  - Mitigation path for pinwheel fabrication issues
    - Stitching errors can be corrected by laser writing tool adjustments
    - Point defects corrected by reduced contamination/larger spot size

- Phase 2 Objectives
  - Develop manufacturable large port-count Rotor switch
    - Replace fiber array with collimator array
    - Replace fiber patch panel w/micro-optics array

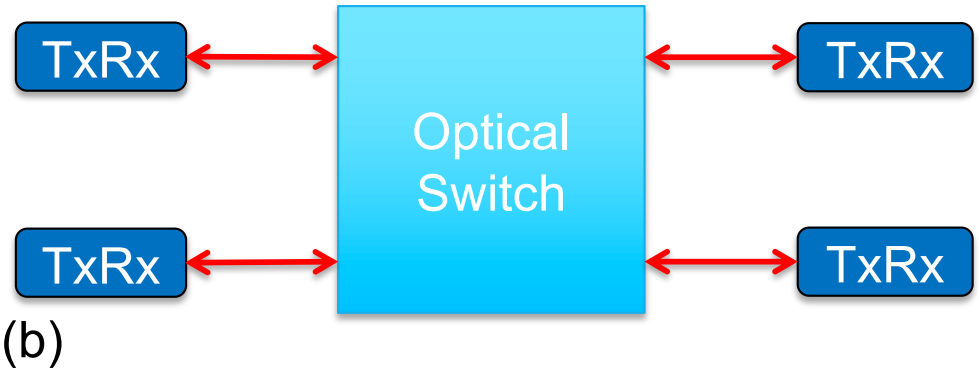
LEED Interconnect objectives

- I. Optically-interconnected, electrical switching



- Switch energy is relatively high
- Link metrics
  - 2 pJ/bit
- BW density
  - 1 Tb/s/cm

- II. Optically switched



- Switch energy is low
- Switch loss is managed w/o amplifiers
  - Link is optimized for margin
  - **Link metrics(1.2)**
    1 pJ/bit excluding laser power
    +1 pJ/bit laser x excess switch loss
    =2 pJ/bit for a lossless switch
  - **Link metric vs ~14 pJ/bit Case I**
- Scales > 100 Tb/s w/WDM

LEED interconnect Phase 2 projects

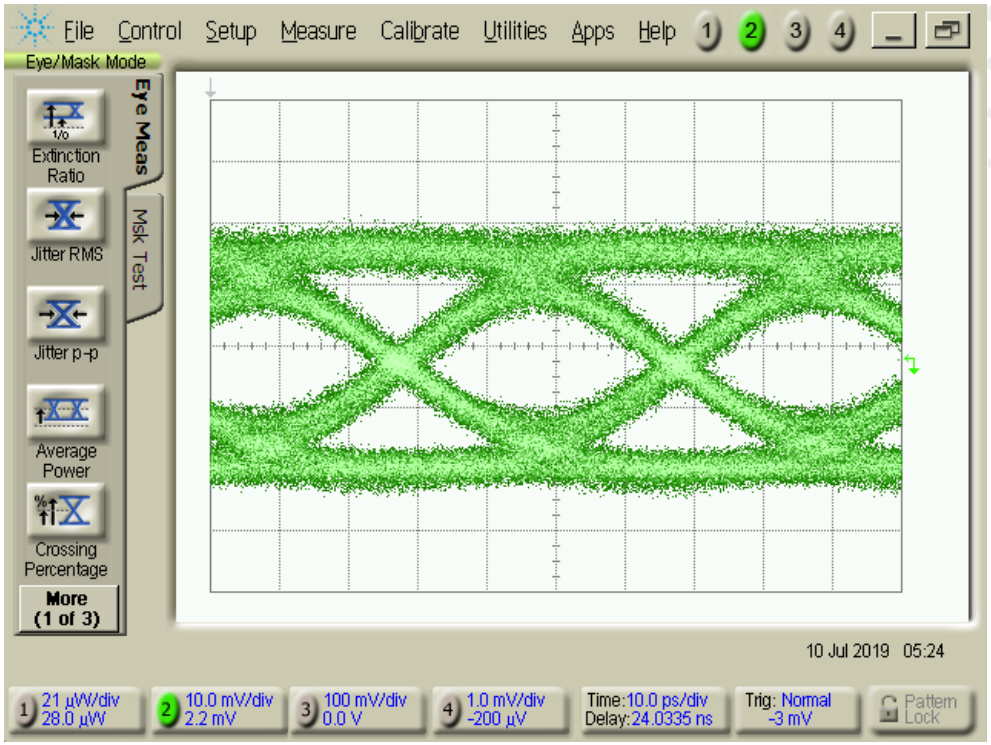
- **Integrated Transmitter**
  - *UCSD: modulator/testing* Forrest Valdez, Shayan Mookerjea
  - *UCSD: laser source testing* Suruj Deka, Shaya Fainmain
  - *Axalume: source array/integration* Y. Ehrlichman, I. G. Yayla, J. Simons, J. E. Cunningham, A. V. Krishnamoorthy

- **Integrated Receiver**
  - *UCSD: demux* Jordon Davis, Shaya Fainman
  - *Axalume: BM Rx front end* same at Tx team
  - *Sandia: demux/APD* Michael Gehl, Doug Trotter, Andrew Starbuck, Tina Dallo, Dana Hood, Andrew Pomerene, Paul Chris DeRose, Tony Lentine

- The rest of the LEED team members all have input to the interconnects (co-design)

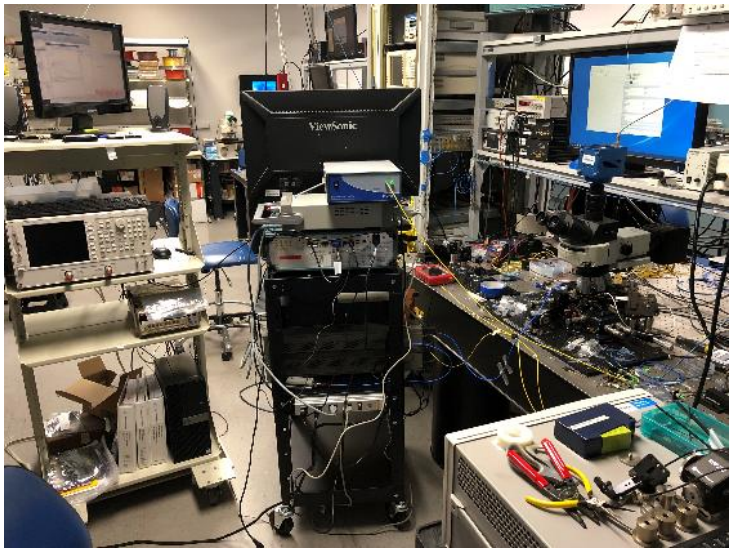
25 Gb/s APDs for integrated BM Rx

- Demonstrate an APD with a responsivity of 10 A/W and a 3 dB bandwidth of 25 Gb/s.
  - Link modeling shows ~ 3.5 A/W is needed to achieve 2 pJ/bit with the switch in the path.
- Results:
  - R=5.4 A/W and B=21.8 GHz, satisfies 25 Gb/s at 2 pJ/bit
  - R=7.1 A/W and B=18.8 GHz, satisfies 25 Gb/s at 2 pJ/bit
  - R=3.1 A/W and B=46.1 GHz, potential for 50 Gb/s
  - R=1.7 A/W and B=71.4 GHz, potential >> 50 Gb/s

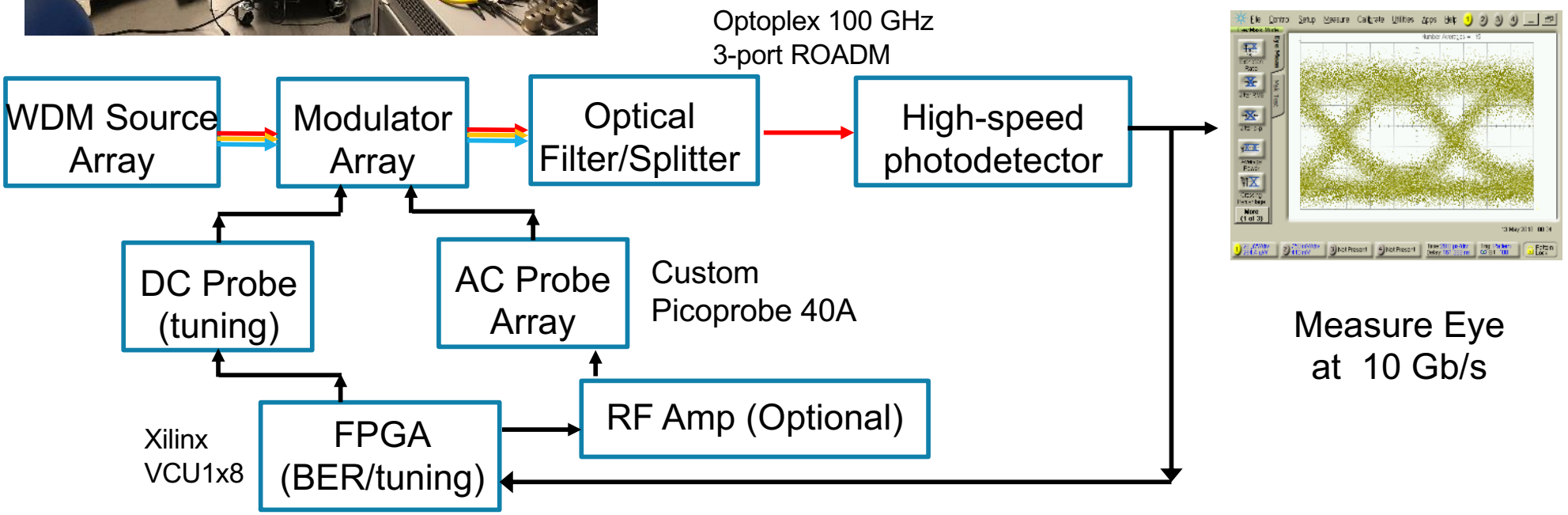


Eye at 25 Gb/s from APD @ 3.5 A/W

High OMA WDM modulator array

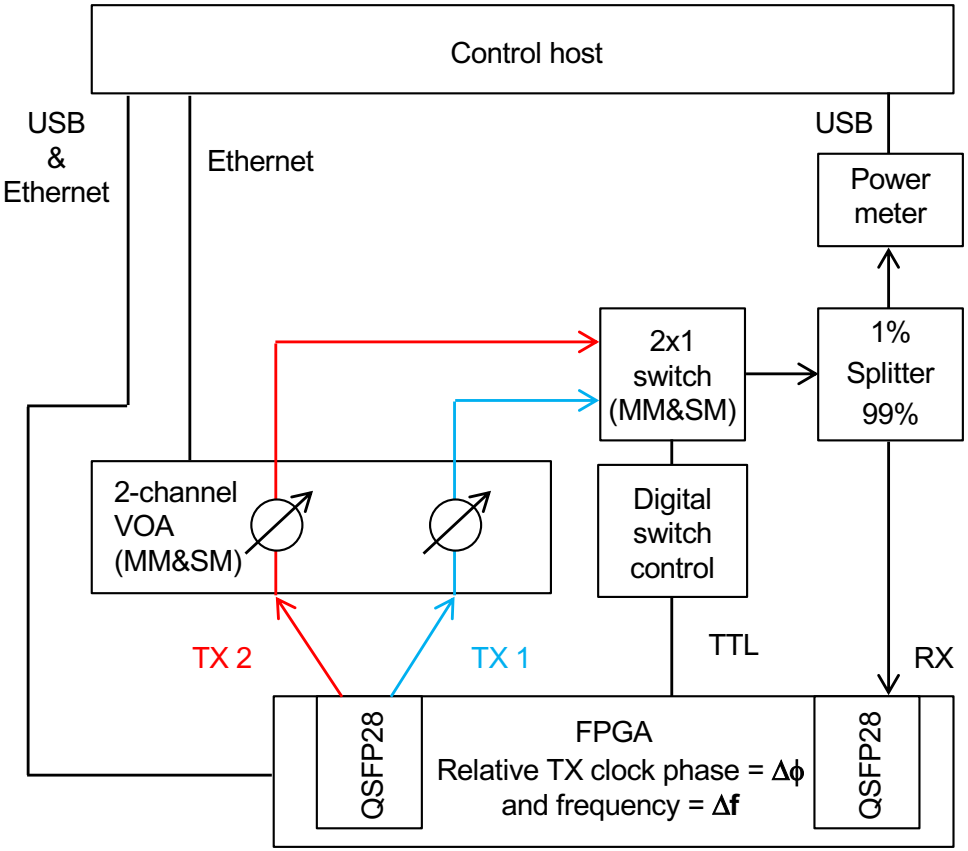


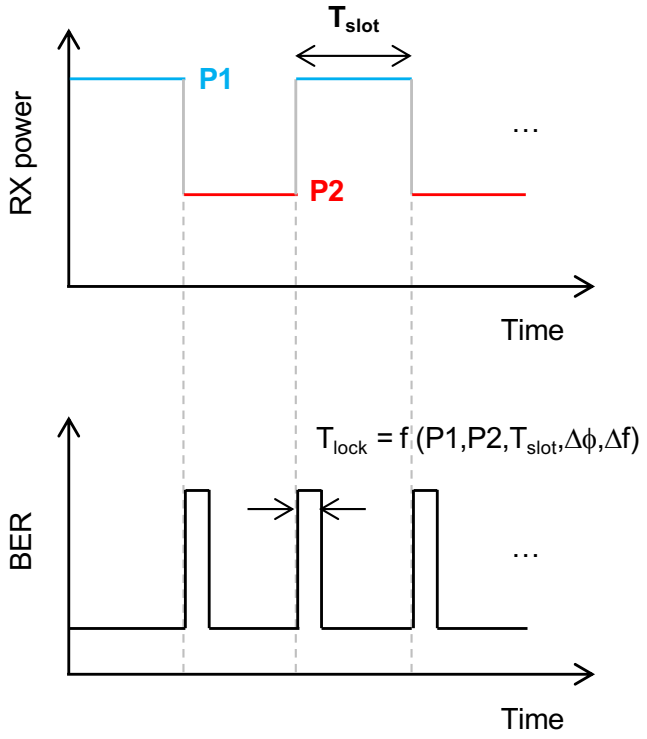
- Goal: Demonstration of an eight channel 25 Gb/s modulator array with comb source and closed loop control
- Challenge: close link w/o optical amp
- Requires detailed device modeling/characterization



Transparent switching testbed

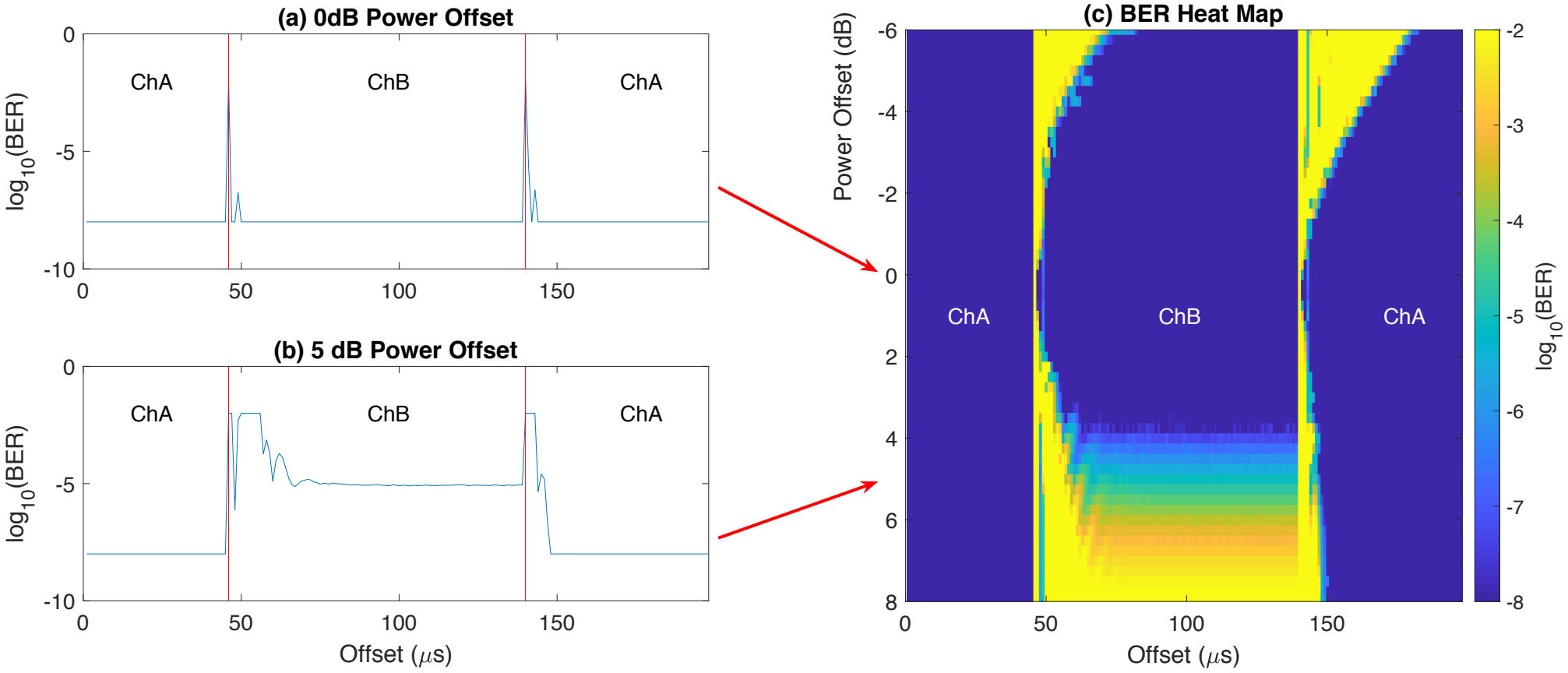
Goal: map the functional relationship of $T_{lock} = f(P1,P2,T_{slot},\Delta\phi,\Delta f)$ for various commercial and LEED-developed transceivers.





Transparent switching testbed operation

Time-resolved BER measurement using two transmitters w/each from a separate PSM4 module



Interconnect: Phase 1 accomplishments/ Phase 2 Objectives

- Phase 1 accomplishments
  - Burst-mode Rx: key feedforward operation demonstrated
  - APD: 25 Gb/s @ 3.5 A/W demonstrated →2 pJ/bit link
  - Modulator Array: Coupled physics/device design process
  - Switched Links: Switched power offset identified as issue

- Phase 2 Objectives
  - Integrated Tx
    - Combines modulator work with source array and control
  - Integrated Rx
    - Combines mux/demux, APD, and BM frontend
  - Goal: Co-optimize photonic Tx/Rx chips to close link w/o amplification & mitigate switched power offset

Conclusions

- Phase 1 of LEED developed the key technologies (network architecture/protocol, switch, and interconnect) for a practical, cost-effective energy-efficient optical network for datacenters and HPC

- Phase 2 will focus on the manufacturable integration of those components to demonstrate a viable use case for energy-efficient optical networking
  - Demonstrate Opera w/viable use cases
  - Manufacturable low-loss large port count Rotor switch
  - Integrated Tx and Rx that can close a link w/o optical amplification & address transient power offset issues